BLIND RT60 ESTIMATION ROBUST ACROSS ROOM SIZES AND SOURCE DISTANCES

Baldwin Dumortier[1,2,3] *and Emmanuel Vincent*[1,2,3]

[1]Inria, Villers-lès-Nancy, F-54600, France
[2]CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3]Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
emmanuel.vincent@inria.fr

ABSTRACT

The reverberation time or RT60 is an essential acoustic parameter of a room. In many situations, the room impulse response (RIR) is not available and the RT60 must be blindly estimated from a speech or music signal. Current methods often implicitly assume that reverberation dominates direct sound, which restricts their applicability to relatively small rooms or distant sound sources. This paper features two contributions. Firstly, we propose a blind RT60 estimation method that is independent of the room size and the source distance by preprocessing the input signal using a beamformer to cancel direct sound and early echoes. Secondly, we perform the largest experimental evaluation to our knowledge using a set of 342 RIRs. We show that the estimation error is significantly reduced even in the case when reverberation dominates.

***Index Terms*—** Reverberation time, blind estimation, spectral decay distribution, direct-to-reverberant ratio

1. INTRODUCTION

The reverberation time or RT60 is one of the main parameters describing the acoustic properties of a room. It is useful to assess the intelligibility of speech and as prior knowledge to perform dereverberation [1, 2], source separation [3] or robust automatic speech recognition (ASR) [4, 5].

The RT60 is defined as the time required for sound to decay by 60 dB once the source has been switched off [6]. It can be calculated from a room impulse response (RIR) using Schroeder's method [7] or approximated from the room characteristics using Sabine's [8] or Eyring's models [9]. In many situations, this information is not available, however, and the RT60 has to be *blindly* estimated from a recorded signal. Current blind estimation methods roughly fall into two categories. Following Polack's RIR model [10], some methods estimate the distribution of the decay rates of the power envelope of the signal over time and map its mode or another statistic to the RT60 via a fixed or a learned mapping [11–16]. Some other methods rely on the quantification of the deformation of cepstral features or modulation features due to reverberation [17, 18]. Three methods were compared in [19] on a set of 3 real-world RIRs and 9 simulated RIRs for a single room and a single source distance to the microphone.

All of these methods explicitly or implicitly assume that reverberation dominates direct sound in the input signal [13] or, in other words, that the direct-to-reverberant ratio (DRR) is below 0 dB. Indeed, the observed decay rates and features exhibit lesser deviation from a clean signal when direct sound dominates. This assumption restricts their applicability to relatively small rooms or distant sound sources. For instance, the critical distance above which the DRR falls below 0 dB is on the order of 57 cm in a small 30 m^3 meeting room with a RT60 of 300 ms [8]. Yet, slightly larger DRRs are still detrimental to source separation or ASR [2]. Robust RT60 estimation across room sizes and source distances is hence necessary.

In this paper, we propose to achieve robust estimation by preprocessing the input signal using a beamformer to enhance reverberation. We evaluate the baseline method in [13] with and without preprocessing on a dataset generated from 342 RIRs, which is the largest set of RIRs considered in this context to our knowledge. The rest of the paper is organized as follows. In Section 2 we briefly introduce Polack's model and the RT60 estimation method in [13]. In Section 3, we discuss the influence of the DRR and we present the proposed preprocessing step. We describe the experimental evaluation in Section 4 and we conclude in Section 5.

2. SPECTRAL DECAY DISTRIBUTION

2.1. RIR model

Let us consider a recorded signal $x(t)$ that is the convolution between a clean source source signal $s(t)$ and the RIR $h(t)$ from the source to the microphone. Assuming that reverberation dominates direct sound, Polack [10] proposed a statistical model for RIRs by which each sample of $h(t)$ is Gaussian distributed with time-dependent variance (or power) $d_h(t)$. The power is exponentially decreasing according to

$$d_h(t) = \sigma^2 e^{\lambda_h t} \tag{1}$$

where σ^2 is a constant and λ_h is a negative *decay rate* that is inversely proportional to the RT60.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2.2. RT60 estimation from the decay distribution

Let us denote by $S(n, f)$ and $X(n, f)$ the short time Fourier transform (STFT) coefficients of the source and the recorded signal in time frame n and frequency bin f. Assuming that the power of $S(n, f)$ locally follows a similar exponential model with decay rate $\lambda_s(n, f)$, the decay rate $\lambda_x(n, f)$ of the power of $X(n, f)$ is given by [13]

$$\lambda_x(n, f) \approx \max[\lambda_h, \lambda_s(n, f)]. \quad (2)$$

This decay rate may be measured via a linear least squares fit of the log-power of $X(n, f)$ on L successive time frames n', $n \leq n' \leq n + N - 1$.

Let us now consider the distribution of λ_x over the time-frequency plane. Remembering that $\lambda_h < 0$, (2) implies that this distribution is similar to the distribution of λ_s for positive values, while it is increased around λ_h for negative values. Wen et al. found the *negative side variance* – that is, the variance of the negative side of the distribution of λ_x – to be well correlated with this increase and they proposed to derive the RT60 from the negative side variance via a second-order polynomial mapping trained on development data. See [13] for details.

This method performed best among the tested methods in [19] in noiseless conditions and it constitutes our baseline in the following.

3. ROBUSTNESS TO THE ROOM SIZE AND THE SOURCE DISTANCE

3.1. Dependency of the decay distribution on the DRR

As mentioned above, Polack's model (1) holds only when reverberation dominates direct sound in $h(t)$. Assuming without loss of generality that direct sound occurs at $t = 0$, the RIR power is better modeled in the general case as [2]

$$d_h(t) = \sigma^2_{\text{dir}}\delta_0(t) + \sigma^2_{\text{rev}}e^{-\lambda_h t} \quad (3)$$

where δ_0 is a Dirac and σ^2_{dir} and σ^2_{rev} represent the power of direct and reverberant sound, respectively. Under this model, the DRR can be computed as

$$\text{DRR} = 10\log_{10}(-\lambda_h\sigma^2_{\text{dir}}/\sigma^2_{\text{rev}}). \quad (4)$$

σ^2_{rev} does not depend on the position in the room, while σ^2_{dir} is inversely proportional to the square of the source distance. Also, σ^2_{rev} decreases with the room size for a given RT60 [9], while σ^2_{dir} does not depend on it. Overall, the DRR can take a large range of negative and positive values for any RT60 and it is smaller for smaller rooms or for distant sources.

The general model (3) allows us to analyze the influence of the DRR on RT60 estimation. When the DRR is large, reverberation is "hidden" by direct sound and the distribution of λ_x becomes identical to that of λ_s for all RT60, so that the RT60 cannot be deduced from it anymore. More crucially, as the DRR increases from negative to positive values, the distribution of λ_x changes continuously from the distribution due to reverberation alone to the one due to direct sound alone. As a consequence, the negative side variance spans a large range of values for a given RT60 which overlaps the values spanned for other RT60s. The relationship between the negative side variance and the RT60 then becomes ambiguous, so that the estimation error increases on average over all DRRs. This observation also holds for the alternative statistics or the feature deformation metrics considered in other methods.

In order to achieve robust estimation with respect to the room size and the source distance, it seems natural to seek to enhance reverberation in the input signal. This goal, which is the opposite of dereverberation, seems rather nontrivial in a single-channel scenario but it is easier to address in a multi-channel scenario by means of beamforming [20]. We consider two different such beamformers below.

3.2. Direct sound removal

From now on, let us assume the availability of two signals $y_1(t)$ and $y_2(t)$ recorded with omnidirectional microphones. Denoting by $Y_1(n, f)$ and $Y_2(n, f)$ the STFT coefficients of $y_1(t)$ and $y_2(t)$, respectively, we subtract $Y_2(n, f)$ from $Y_1(n, f)$ with a suitable complex-valued weight $w(f)$ so as to yield a single-channel signal $X(n, f)$:

$$X(n, f) = Y_1(n, f) - w(f)\, Y_2(n, f). \quad (5)$$

The corresponding time-domain signal $x(t)$ is then obtained by inverse STFT and it is used as input to the RT60 estimation method in Section 2. Note that equalization of the beamformer output does not matter here, as the decay rates $\lambda_x(n, f)$ are separately estimated in each frequency bin.

The first proposed beamformer aims to remove direct sound by steering a null in the source direction. We estimate the time difference of arrival (TDOA) τ in samples between the two microphones using the variant in [21] of the generalized cross-correlation with phase transform (GCC-PHAT) method implemented in the BSS Locate toolbox[1], which was found to work best in the evaluation in [22]. We also estimate the intensity ratio between the two microphones as $r = \sum_{nf} |Y_1(n, f)|^2 / \sum_{nf} |Y_2(n, f)|^2$. The weight $w(f)$ is then derived as

$$w(f) = \sqrt{r}\, e^{2i\pi\tau f/F} \quad (6)$$

with F denoting the FFT size.

3.3. Direct sound and early echoes removal

The second proposed beamformer aims to remove both direct sound and early echoes. Assuming that the correlation between the late reverberant parts of $y_1(t)$ and $y_2(t)$ is smaller

[1] http://bass-db.gforge.inria.fr/bss_locate/

than between their direct and early parts, this is achieved by choosing $w(f)$ so as to decorrelate $Y_1(n,f)$ from $Y_2(n,f)$:

$$w(f) = \frac{\sum_n Y_1(n,f)Y_2(n,f)^*}{\sum_n |Y_2(n,f)|^2} \quad (7)$$

with * denoting complex conjugation.

4. EXPERIMENTAL EVALUATION

4.1. Data and algorithm settings

We evaluated the benefit of the above two beamformers for the baseline method in Section 2. Motivated by an eventual application to the CHiME Challenge scenario [23], we generated reverberant signals by convolving anechoic continuous speech from the Grid corpus [24] with RIRs simulated via the source image method [25] using Roomsimove[2]. This procedure makes it possible to evaluate a much larger number of room sizes and source distances than current real-world RIR datasets and it was shown to yield comparable performance to real-world RIRs and real-world recordings in [13, 19].

For each of 34 speakers, we concatenated several utterances into a 1 min signal. We generated 342 RIRs for two microphones spaced by 16 cm with the following settings (discarding the settings for which the source is outside the room):

- 9 target RT60s from 0.2 s to 1 s in 0.1 s steps,
- 3 room sizes: $1.92 \times 1.92 \times 1.82$ m (small), $3.85 \times 3.85 \times 3.65$ m (medium), and $7.7 \times 7.7 \times 7.3$ m (large),
- 3 source distances: 0.1 m, 0.5 m, and 2.5 m,
- 5 source directions: $-\pi/4$, $-\pi/8$, 0, $\pi/8$, and $\pi/4$.

Simulation was conducted using absorption coefficients computed via Eyring's approximation [9], so that the actual RT60 slightly differed from the target RT60. The actual RT60 was measured using Schroeder's backward integration method [7].

Half of the clean speech signals and the RIRs were used for training and the other half for testing. For each reverberant signal, the DRR was computed as the energy ratio of the direct and the reverberant part of $x(t)$.

All signals were sampled at 16 kHz. The STFT was computed using 1024-sample Hann windows with 50% overlap for beamforming, and 256-sample Hamming windows with 75% overlap for RT60 estimation as in [13]. The decay rates were measured on $N = 20$ successive time frames (92 ms).

4.2. Distribution of the negative side variance

Figure 1 depicts the negative side variances measured on the unprocessed training set. The values for different RT60s significantly overlap. For instance, a negative side variance of

[2]http://www.loria.fr/~evincent/Roomsimove.zip

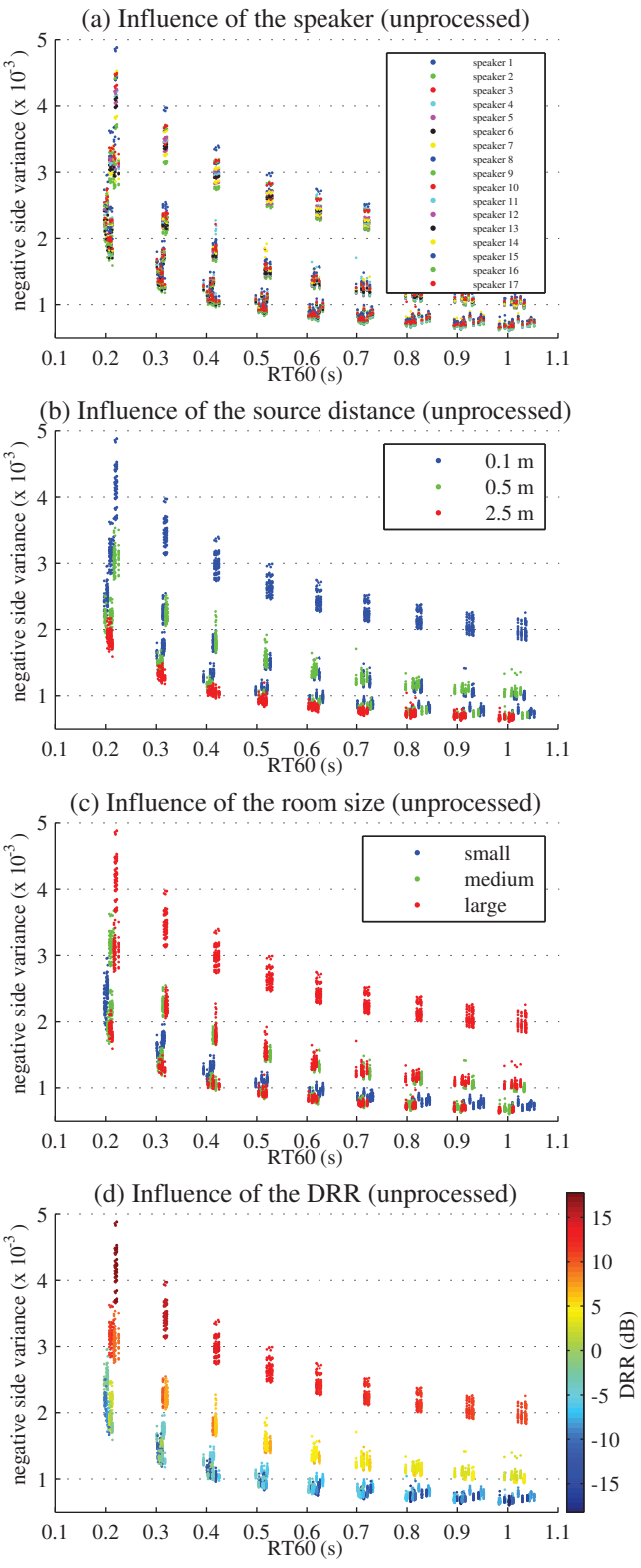

Fig. 1. Negative side variance as a function of the true RT60 on the unprocessed training set (one point per sample). Colors show the influence of (a) the speaker, (b) the source-to-microphone distance, (c) the room size, and (d) the DRR.

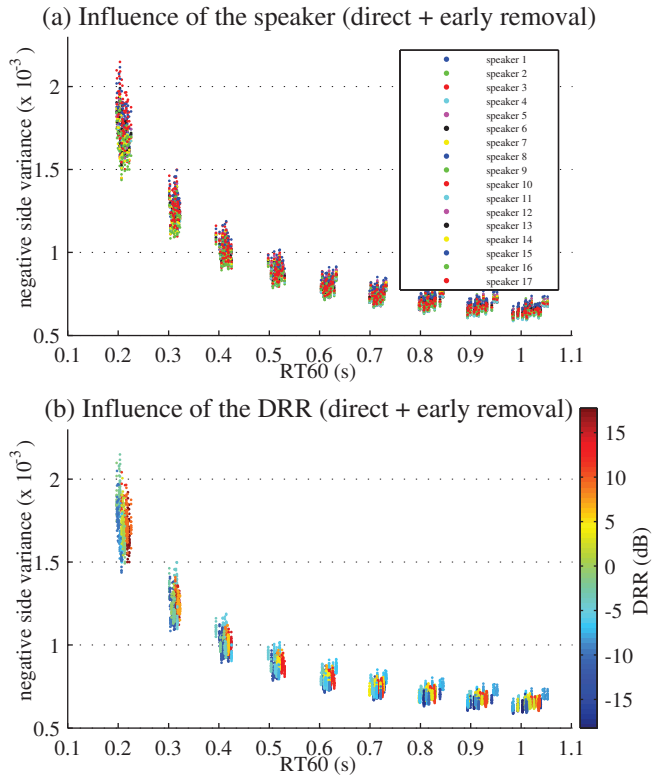

Fig. 2. Negative side variance as a function of the true RT60 on the training set after direct sound and early echoes removal (one point per sample). Colors show the influence of (a) the speaker, (b) the DRR.

2×10^{-3} may correspond to any RT60. The measured value depends on the speaker, but not as much as on the room size and on the source distance. Large values typically correspond to large rooms or close sources, while small values do not imply any particular size or distance. For a given RT60, the DRR turns out to be best correlated with the measured value, with large DRRs corresponding to large negative side variances and vice-versa.

The picture changes after removing direct sound and early echoes, as shown in Figure 2. The negative side variances exhibit smaller overlap and they become independent of the DRR. The residual variability can be attributed to the speaker (e.g., large values for the blue speaker and small values for the green speaker). Indeed, the spectral decay distribution of each speaker is intrinsically related to his/her rate of speech.

4.3. RT60 estimation error

Table 1 shows the relative RT60 estimation error in percent as used in [13]. The worst-case and the root mean square (RMS) error over all test samples are reported for both the full test set and the subset with negative DRR, that is the intended field of application of the original method in [13].

Test set	Preprocessing	max RE	RMS RE
full	none [13]	98%	34%
	direct sound removal	77%	21%
	direct + early removal	**34%**	**10%**
DRR < 0 dB	none [13]	52%	13%
	direct sound removal	54%	13%
	direct + early removal	**33%**	**10%**

Table 1. Worst-case (max RE) and root mean square (RMS RE) relative estimation error with and without preprocessing.

Without preprocessing, the error is as large as 98% worst-case on the full set. The error achieved on the subset with negative DRR is smaller but still larger than reported in previous evaluations with RIRs simulated for a single room and a single source distance to the microphone in [13] (on the order of 15% worst-case) and in [19, Fig. 1] (on the order of 35% worst-case and 10% RMS).

The proposed beamformers both improve performance on the full set. The improvement is greater for the second beamformer which removes direct sound and early echoes than for the first beamformer which removes direct sound only. Crucially, the second beamformer also improves performance on the subset with negative DRR and it achieves the same error on the full set than on this subset, namely 33-34% worst-case and 10% RMS. Together with Figure 2, this confirms that the resulting RT60 estimation method is truly independent of the room size, the source distance, and the DRR, and that the remaining error can be attributed to speaker variability.

5. CONCLUSION

We proposed to enhance reverberation as a preprocessing step for blind RT60 estimation. We tested two alternative beamformers on a large dataset generated from 342 RIRs. The results showed that the beamformer removing both direct sound and early echoes performed best and that the resulting RT60 estimates become truly robust to the room size and the source distance. This makes it possible to use this estimator in any scenario, including in scenarios with positive DRR which are still critical for source separation or ASR [2]. In the future, we aim to expand this study to other early reverberation suppression techniques [26], other RT60 estimation techniques [14,16–18], other microphone distances, and real-world noise scenarios such as in the CHiME Challenge [23]. This implies to remove as much as possible the direct sound and the early echoes of all sources (speech and noise), which may be an easier task than separating and dereverberating them.

6. ACKNOWLEDGMENT

This work was supported by the EUREKA Eurostars i3Dmusic project funded by Bpifrance.

7. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[2] E. A. P. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.

[3] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Spatial location priors for Gaussian model based reverberant audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2013, pp. 149, 2013.

[4] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[5] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1692–1707, 2010.

[6] International Standards Organization, "3382. Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters," 1997.

[7] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, pp. 409–412, 1965.

[8] W. C. Sabine and M. D. Egan, "Collected papers on acoustics," *The Journal of the Acoustical Society of America*, vol. 95, no. 6, pp. 3679–3680, 1994.

[9] C. F. Eyring, "Reverberation time in dead rooms," *The Journal of the Acoustical Society of America*, vol. 1, no. 2A, pp. 217–241, 1930.

[10] J.-D. Polack, *La transmission de l'energie sonore dans les salles*, Ph.D. thesis, Université du Maine, 1988.

[11] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O Brien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[12] A. Baskind and A. de Cheveigne, "Pitch-tracking of reverberant sounds, application to spatial description of sound scenes," in *Proc. AES 24th Int. Conf.*, 2003, Paper number 34.

[13] J. Y.-C. Wen, E. A. P. Habets, and P. A. Naylor, "Blind estimation of reverberation time based on the distribution of signal decay rates," in *Proc. ICASSP*, 2008, pp. 329–332.

[14] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. IWAENC*, 2010.

[15] N. López, Y. Grenier, and I. Bourmeyster, "Low variance blind estimation of the reverberation time," in *Proc. IWAENC*, 2012.

[16] R. Talmon and E. A. P. Habets, "Blind reverberation time estimation by intrinsic modeling of reverberant speech," in *Proc. ICASSP*, 2013, pp. 156–160.

[17] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[18] F. Xiong, S. Goetze, and B. T. Meyer, "Blind estimation of reverberation time based on spectro-temporal modulation filtering," in *Proc. ICASSP*, 2013, pp. 443–447.

[19] N. D. Gaubitch, H. W. Löllmann, M. Jeub, T. H. Falk, P. A. Naylor, P. Vary, and M. Brookes, "Performance comparison of algorithms for blind reverberation time estimation from speech," in *Proc. IWAENC*, 2012.

[20] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, 2001.

[21] B. Loesch and B. Yang, "Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions," in *Proc. LVA/ICA*, 2010, pp. 41–48.

[22] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[23] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Proc. ASRU*, 2013.

[24] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Journal of the Acoustical Society of America*, vol. 120, pp. 2421–2424, 2006.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[26] A. Schwarz, K. Reindl, and W. Kellermann, "A two-channel reverberation suppression scheme based on blind signal separation and Wiener filtering," in *Proc. ICASSP*, 2012, pp. 113–116.